Docket No: VDX-5002 CIP

COLORECTAL CANCER PROGNOSTICS

BACKGROUND

This application is a continuation-in-part of US Application No. 10/651,237 filed August 28, 2003

This invention relates to prognostics for colorectal cancer based on the gene expression profiles of biological samples.

Colorectal cancer is a heterogenous disease with complex origins. Once a patient is treated for colorectal cancer, the likelihood of a recurrence is related to the degree of tumor penetration through the bowel wall and the presence or absence of nodal involvement. These characteristics are the basis for the current staging system defined by Duke's classification. Duke's A disease is confined to submucosa layers of colon or rectum. Duke's B tumor invades through muscularis propria and could penetrate the wall of colon or rectum. Duke's C disease includes any degree of bowel wall invasion with regional lymph node metastasis.

Surgical resection is highly effective for early stage colorectal cancers, providing cure rates of 95% in Duke's A and 75% in Duke's B patients. The presence of positive lymph node in Duke's C disease predicts a 60% likelihood of recurrence within five years. Treatment of Duke's C patients with a post surgical course of chemotherapy reduces the recurrence rate to 40%-50%, and is now the standard of care for Duke's C patients. Because of the relatively low rate of reoccurrence, the benefit of post surgical chemotherapy in Duke' B has been harder to detect and remains controversial. However, the Duke's B classification is imperfect as approximately 20 – 30% of these patients behave more like Duke's C and relapse within a 5-year timeframe.

There is clearly a need to identify better prognostic factors than nodal involvement for guiding selection of Duke's B into those that are likely to relapse and those that will survive. In commonly owned US Patent Application 10/403,499 to

Wang, gene expression profiles prognostic for colon cancer were presented. This specification presents different gene expression profiles.

SUMMARY OF THE INVENTION

The invention is a method of assessing the likelihood of a recurrence of colorectal cancer in a patient diagnosed with or treated for colorectal cancer. The method involves the analysis of a gene expression profile.

In one aspect of the invention, the gene expression profile includes at least seven particular genes.

In another aspect of the invention, the gene expression profile includes at least fifteen particular genes.

In yet another aspect of the invention, the gene expression profile includes the seven particular genes as well as the fifteen particular genes described above. In one embodiment, the gene profile comprises twenty-three genes.

Articles used in practicing the methods are also an aspect of the invention. Such articles include gene expression profiles or representations of them that are fixed in machine-readable media such as computer readable media.

Articles used to identify gene expression profiles can also include substrates or surfaces, such as microarrays, to capture and/or indicate the presence, absence, or degree of gene expression.

In yet another aspect of the invention, kits include reagents for conducting the gene expression analysis prognostic of colorectal caner recurrence.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a standard Kaplan-Meier Plot constructed from the independent patient data set of 27 patients (14 survivors, 13 relapses) as described in the Examples for the analysis of the seven gene portfolio. Two classes of patients are indicated as predicted by chip data. The vertical axis shows the probability of disease-free survival among patients in each class.

Fig. 2 is a standard Kaplan-Meier Plot constructed from the independent patient data set of 9 patients (6 survivors, 3 relapses) as described in the Examples for the analysis of the 15 gene portfolio. Two classes of patients are indicated as predicted by chip data. The vertical axis shows the probability of disease-free survival among patients in each class.

Fig. 3 is a standard Kaplan-Meier Plot constructed from patient data as described in the Examples and using the 22- gene profile with the inclusion of Cadherin 17 (Seq. ID 6) to the portfolio. Thirty-six samples were tested (20 survivor, 16 relapses) Two classes of patients are indicated as predicted by chip data of the 23-gene panel. The vertical axis shows the probability of disease-free survival among patients in each class.

DETAILED DESCRIPTION

The mere presence or absence of particular nucleic acid sequences in a tissue sample has only rarely been found to have diagnostic or prognostic value. Information about the expression of various proteins, peptides or mRNA, on the other hand, is increasingly viewed as important. The mere presence of nucleic acid sequences having the potential to express proteins, peptides, or mRNA (such sequences referred to as "genes") within the genome by itself is not determinative of whether a protein, peptide, or mRNA is expressed in a given cell. Whether or not a given gene capable of expressing proteins, peptides, or mRNA does so and to what extent such expression occurs, if at all, is determined by a variety of complex factors. Irrespective of difficulties in understanding and assessing these factors, assaying gene expression can provide useful information about the occurrence of important events such as tumerogenesis, metastasis, apoptosis, and other clinically relevant phenomena. Relative indications of the degree to which genes are active or inactive can be found in gene expression profiles. The gene expression profiles of this invention are used to provide a prognosis and treat patients for colorectal cancer.

Sample preparation requires the collection of patient samples. Patient samples used in the inventive method are those that are suspected of containing diseased cells such as epithelial cells taken from the primary tumor in a colon sample or from surgical margins. Laser Capture Microdisection (LCM*)* technology is one way to select the cells to be studied, minimizing variability caused by cell type heterogeneity. Consequently, moderate or small changes in gene expression between normal and cancerous cells can be readily detected. Samples can also comprise circulating epithelial cells extracted from peripheral blood. These can be obtained according to a number of methods but the most preferred method is the magnetic separation technique described in U.S. Patent 6,136,182 assigned to Immunivest Corp which is incorporated herein by reference. Once the sample containing the cells of interest has been obtained, RNA is extracted and amplified and a gene expression profile is obtained, preferably via micro-array, for genes in the appropriate portfolios.

Preferred methods for establishing gene expression profiles include determining the amount of RNA that is produced by a gene that can code for a protein or peptide. This is accomplished by reverse transcriptase PCR (RT-PCR), competitive RT-PCR, real time RT-PCR, differential display RT-PCR, Northern Blot analysis and other related tests. While it is possible to conduct these techniques using individual PCR reactions, it is best to amplify complimentary DNA (cDNA) or complimentary RNA (cRNA) produced from mRNA and analyze it via microarray. A number of different array configurations and methods for their production are known to those of skill in the art and are described in U.S. Patents such as: 5,445,934; 5,532,128; 5,556,752; 5,242,974; 5,384,261; 5,405,783; 5,412,087; 5,424,186; 5,429,807; 5,436,327; 5,472,672; 5,527,681; 5,529,756; 5,545,531; 5,554,501; 5,561,071; 5,571,639; 5,593,839; 5,599,695; 5,624,711; 5,658,734; and 5,700,637; the disclosures of which are incorporated herein by reference.

Microarray technology allows for the measurement of the steady-state mRNA level of thousands of genes simultaneously thereby presenting a powerful tool for identifying effects such as the onset, arrest, or modulation of uncontrolled cell proliferation. Two microarray technologies are currently in wide use. The first are cDNA arrays and the second are oligonucleotide arrays. Although differences exist in the construction of these chips, essentially all downstream data analysis and output are the same. The product of these analyses are typically measurements of the intensity of the signal received from a labeled probe used to detect a cDNA sequence from the sample that hybridizes to a nucleic acid sequence at a known location on the microarray. Typically, the intensity of the signal is proportional to the quantity of cDNA, and thus mRNA, expressed in the sample cells. A large number of such techniques are available and useful. Preferred methods for determining gene expression can be found in US Patents 6,271,002 to Linsley, et al.; 6,218,122 to Friend, et al.; 6,218,114 to Peck, et al.; and 6,004,755 to Wang, et al., the disclosure of each of which is incorporated herein by reference.

Analysis of the expression levels is conducted by comparing such signal intensities. This is best done by generating a ratio matrix of the expression intensities of genes in a test sample versus those in a control sample. For instance, the gene expression intensities from a diseased tissue can be compared with the expression intensities generated from normal tissue of the same type (e.g., diseased colon tissue sample vs. normal colon tissue sample). A ratio of these expression intensities indicates the fold-change in gene expression between the test and control samples.

Gene expression profiles can also be displayed in a number of ways. The most common method is to arrange a raw fluorescence intensities or ratio matrix into a graphical dendogram where columns indicate test samples and rows indicate genes. The data is arranged so genes that have similar expression profiles are proximal to each other. The expression ratio for each gene is visualized as a color. For example, a ratio less than one (indicating down-regulation) may appear in the blue portion of the spectrum while a ratio greater than one (indicating up-regulation) may appear as a color in the red portion of the spectrum. Commercially available computer software programs are available to display such data including "GENESPRING" from Silicon Genetics, Inc. and "DISCOVERY" and "INFER" software from Partek, Inc.

Modulated genes used in the methods of the invention are described in the Examples. The genes that are differentially expressed are either up regulated or down regulated in patients with a relapse of colon cancer relative to those without a relapse. Up regulation and down regulation are relative terms meaning that a detectable difference (beyond the contribution of noise in the system used to measure it) is found in the amount of expression of the genes relative to some baseline. In this case, the baseline is the measured gene expression of a non-relapsing patient. The genes of interest in the diseased cells (from the relapsing patients) are then either up regulated or down regulated relative to the baseline level using the same measurement method. Diseased, in this context, refers to an alteration of the state of a body that interrupts or disturbs, or has the potential to disturb, proper performance of bodily functions as occurs with the uncontrolled proliferation of cells. Someone is diagnosed with a disease when some aspect of that person's genotype or phenotype is consistent with the presence of the disease. However, the act of conducting a diagnosis or prognosis includes the determination of disease/status issues such as determining the likelihood of relapse and therapy monitoring. In therapy monitoring, clinical judgments are made regarding the effect of a given course of therapy by comparing the expression of genes over time to determine whether the gene expression profiles have changed or are changing to patterns more consistent with normal tissue.

Preferably, levels of up and down regulation are distinguished based on fold changes of the intensity measurements of hybridized microarray probes. A 2.0 fold difference is preferred for making such distinctions or a p-value less than .05. That is, before a gene is said to be differentially expressed in diseased/relapsing versus normal/non-relapsing cells, the diseased cell is found to yield at least 2 more, or 2 times less intensity than the normal cells. The greater the fold difference, the more preferred is use of the gene as a diagnostic or prognostic tool. Genes selected for the gene expression profiles of the instant invention have expression levels that result in the generation of a signal that is distinguishable from those of the normal or non-modulated genes by an amount that exceeds background using clinical laboratory instrumentation.

Statistical values can be used to confidently distinguish modulated from non-modulated genes and noise. Statistical tests find the genes most significantly different between diverse groups of samples. The Student's t-test is an example of a robust statistical test that can be used to find significant differences between two groups. The lower the p-value, the more compelling the evidence that the gene is showing a difference between the different groups. Nevertheless, since microarrays measure more than one gene at a time, tens of thousands of statistical tests may be asked at one time. Because of this, one is unlikely to see small p-values just by chance and adjustments for this using a Sidak correction as well as a randomization/permutation experiment can be made. A p-value less than .05 by the t-test is evidence that the gene is significantly different. More compelling evidence is a p-value less then .05 after the Sidak correction is factored in. For a large number of samples in each group, a p-value less than 0.05 after the randomization/permutation test is the most compelling evidence of a significant difference.

Another parameter that can be used to select genes that generate a signal that is greater than that of the non-modulated gene or noise is the use of a measurement of absolute signal difference. Preferably, the signal generated by the modulated gene expression is at least 20% different than those of the normal or non-modulated gene (on an absolute basis). It is even more preferred that such genes produce expression patterns that are at least 30% different than those of normal or non-modulated genes.

Genes can be grouped so that information obtained about the set of genes in the group provides a sound basis for making a clinically relevant judgment such as a diagnosis, prognosis, or treatment choice. These sets of genes make up the portfolios of the invention. In this case, the judgments supported by the portfolios involve colorectal cancer and its chance of recurrence, most preferably, among Dukes B patients. As with most diagnostic markers, it is often desirable to use the fewest number of markers sufficient to make a correct medical judgment. This prevents a delay in treatment pending further analysis as well inappropriate use of time and resources.

Preferably, portfolios are established such that the combination of genes in the portfolio exhibit improved sensitivity and specificity relative to individual genes or randomly selected combinations of genes. In the context of the instant invention, the sensitivity of the portfolio can be reflected in the fold differences exhibited by a gene's expression in the diseased state relative to the normal state. Specificity can be reflected in statistical measurements of the correlation of the signaling of gene expression with the condition of interest. For example, standard deviation can be a used as such a measurement. In considering a group of genes for inclusion in a portfolio, a small standard deviation in expression measurements correlates with greater specificity. Other measurements of variation such as correlation coefficients can also be used in this capacity.

One method of establishing gene expression portfolios is through the use of optimization algorithms such as the mean variance algorithm widely used in establishing stock portfolios. This method is described in detail in the patent application entitled "Portfolio Selection" by Tim Jatkoe, et. al., filed on March 21, 2003. Essentially, the method calls for the establishment of a set of inputs (stocks in financial applications, expression as measured by intensity here) that will optimize the return (e.g., signal that is generated) one receives for using it while minimizing the variability of the return. Many commercial software programs are available to conduct such operations. "Wagner Associates Mean-Variance Optimization Application", referred to as "Wagner Software" throughout this specification, is preferred. This software uses functions from the "Wagner Associates Mean-Variance Optimization Library" to determine an efficient frontier and optimal portfolios in the Markowitz sense.is preferred. Use of this type of software requires that microarray data be transformed so that it can be treated as an input in the way stock return and risk measurements are used when the software is used for its intended financial analysis purposes.

The process of selecting a portfolio can also include the application of heuristic rules. Preferably, such rules are formulated based on biology and an understanding of the technology used to produce clinical results. More preferably, they are applied to output from the optimization method. For example, the mean variance method of portfolio selection can be applied to microarray data for a number of genes differentially expressed in subjects with colorectal cancer. Output from the method would be an optimized set of genes that could include some genes that are expressed in peripheral blood as well as in diseased tissue. If samples used in the testing method are obtained from peripheral blood and certain genes differentially expressed in instances of breast cancer could also be differentially expressed in peripheral blood, then a heuristic rule can be applied in which a portfolio is selected from the efficient frontier excluding those that are differentially expressed in peripheral blood. Of course, the rule can be applied prior to the formation of the efficient frontier by, for example, applying the rule during data pre-selection.

Other heuristic rules can be applied that are not necessarily related to the biology in question. For example, one can apply a rule that only a certain percentage of the portfolio can be represented by a particular gene or group of genes. Commercially available software such as the Wagner Software readily accommodates these types of heuristics. This can be useful, for example, when factors other than accuracy and precision (e.g., anticipated licensing fees) have an impact on the desirability of including one or more genes.

One method of the invention involves comparing gene expression profiles for various genes (or portfolios) to ascribe prognoses. The gene expression profiles of each of the genes comprising the portfolio are fixed in a medium such as a computer readable medium. This can take a number of forms. For example, a table can be established into which the range of signals (e.g., intensity measurements) indicative of disease is input. Actual patient data can then be compared to the values in the table to determine whether the patient samples are normal or diseased. In a more sophisticated embodiment, patterns of the expression signals (e.g., flourescent intensity) are recorded digitally or graphically. The gene expression patterns from the gene portfolios used in conjunction with patient samples are then compared to the expression patterns. Pattern comparison software can then be used to determine whether the patient samples have a pattern indicative of recurrence of the disease. Of course, these comparisons can also be used to determine whether the patient is not likely to experience disease recurrence. The expression profiles of the samples are then compared to the portfolio of a control cell. If the sample expression patterns are consistent with the expression pattern for recurrence of a colorectal cancer then (in the absence of countervailing medical considerations) the patient is treated as one would treat a relapse patient. If the sample expression patterns are consistent with the expression pattern from the normal/control cell then the patient is diagnosed negative for colorectal cancer.

The preferred profiles of this invention are the seven-gene portfolio shown in Table 2 and the fifteen-gene portfolio shown in Table 3. It is more preferred to use a portfolio in which both seven and fifteen gene groups are combined. Gene expression portfolios made up another independently verified colorectal prognostic gene such as Cadherin 17 (Seq. ID No. 6) together with the combination of genes in both Table 2 and Table 3 are most preferred (Table 4). This most preferred portfolio best segregates Duke's B patients at high risk of relapse from those who are not. Once the high-risk patients are identified they can then be treated with adjuvant therapy. Other independently verified prognostic genes that can be used in place of Cadherin 17 include, without limitation, genes that correspond to Seq ID No. 29-94.

In this invention, the most preferred method for analyzing the gene expression pattern of a patient to determine prognosis of colon cancer is through the use of a Cox hazard analysis program. Most preferably, the analysis is conducted using S-Plus software (commercially available from Insightful Corporation). Using such methods, a gene expression profile is compared to that of a profile that confidently represents relapse (i.e., expression levels for the combination of genes in the profile is indicative of relapse). The Cox hazard model with the established threshold is used to compare the similarity of the two profiles (known relapse versus patient) and then determines whether the patient profile exceeds the threshold. If it does, then the patient is classified as one who will relapse and is accorded treatment such as adjuvant therapy. If the patient profile does not exceed the threshold then they are classified as a non-relapsing patient. Other analytical tools can also be used to answer the same question such as, linear discriminate analysis, logistic regression and neural network approaches.

Numerous other well-known methods of pattern recognition are available. The following references provide some examples:

Weighted Voting:

Golub, TR., Slonim, DK., Tamaya, P., Huard, C., Gaasenbeek, M., Mesirov, JP., Coller, H., Loh, L., Downing, JR., Caligiuri, MA., Bloomfield, CD., Lander, ES. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science 286:531-537, 1999

Support Vector Machines:

Su, AI., Welsh, JB., Sapinoso, LM., Kern, SG., Dimitrov, P., Lapp, H., Schultz, PG., Powell, SM., Moskaluk, CA., Frierson, HF. Jr., Hampton, GM. *Molecular classification of human carcinomas by use of gene expression signatures.* Cancer Research 61:7388-93, 2001

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, CH., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, JP., Poggio, T., Gerald, W., Loda, M., Lander, ES., Gould, TR. *Multiclass cancer diagnosis using tumor gene expression signatures* Proceedings of the National Academy of Sciences of the USA 98:15149-15154, 2001

K-nearest Neighbors:

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, CH., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, JP., Poggio, T., Gerald, W., Loda, M., Lander, ES., Gould, TR. *Multiclass cancer diagnosis using tumor gene expression signatures* Proceedings of the National Academy of Sciences of the USA 98:15149-15154, 2001

Correlation Coefficients:

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH.Gene expression profiling predicts clinical outcome of breast cancer.
Nature. 2002 Jan 31;415(6871):530-6.

The gene expression profiles of this invention can also be used in conjunction with other non-genetic diagnostic methods useful in cancer diagnosis, prognosis, or treatment monitoring. For example, in some circumstances it is beneficial to combine the diagnostic power of the gene expression based methods described above with data from conventional markers such as serum protein markers (e.g., carcinoembryonic antigen). A range of such markers exists including such analytes as CEA. In one such method, blood is periodically taken from a treated patient and then subjected to an enzyme immunoassay for one of the serum markers described above. When the concentration of the marker suggests the return of tumors or failure of therapy, a sample source amenable to gene expression analysis is taken. Where a suspicious mass exists, a fine needle aspirate is taken and gene expression profiles of cells taken from the mass are then analyzed as described above. Alternatively, tissue samples may be taken from areas adjacent to the tissue from which a tumor was previously removed. This approach can be particularly useful when other testing produces ambiguous results.

Articles of this invention include representations of the gene expression profiles useful for treating, diagnosing, prognosticating, and otherwise assessing diseases. These profile representations are reduced to a medium that can be automatically read by a machine such as computer readable media (magnetic, optical, and the like). The articles can also include instructions for assessing the gene expression profiles in such media. For example, the articles may comprise a CD ROM having computer instructions for comparing gene expression profiles of the portfolios of genes described above. The articles may also have gene expression profiles digitally recorded therein so that they may be compared with gene expression data from patient samples. Alternatively, the profiles can be recorded in different representational format. A graphical recordation is one such format. Clustering algorithms such as those incorporated in "DISCOVERY" and "INFER" software from Partek, Inc. mentioned above can best assist in the visualization of such data.

Different types of articles of manufacture according to the invention are media or formatted assays used to reveal gene expression profiles. These can comprise, for example, microarrays in which sequence complements or probes are affixed to a matrix to which the sequences indicative of the genes of interest combine creating a readable determinant of their presence. Alternatively, articles according to the invention can be fashioned into reagent kits for conducting hybridization, amplification, and signal generation indicative of the level of expression of the genes of interest for detecting colorectal cancer.

Kits made according to the invention include formatted assays for determining the gene expression profiles. These can include all or some of the materials needed to conduct the assays such as reagents and instructions.

The invention is further illustrated by the following non-limiting examples.

Examples: Genes analyzed according to this invention are typically related to full-length nucleic acid sequences that code for the production of a protein or peptide. One skilled in the art will recognize that identification of full-length sequences is not necessary from an analytical point of view. That is, portions of the sequences or ESTs can be selected according to well-known principles for which probes can be designed to assess gene expression for the corresponding gene.

Example 1- Sample Handling and LCM.

Fresh frozen tissue samples were collected from patients who had surgery for colorectal tumors. The samples that were used were from 63 patients staged with Duke's B according to standard clinical diagnostics and pathology. Clinical outcome of the patients was known. Thirty-six of the patients have remained disease-free for more than 3 years while 27 patients had tumor relapse within 3 years.

The tissues were snap frozen in liquid nitrogen within 20-30 minutes of harvesting, and stored at –80C° thereafter. For laser capture, the samples were cut (6μm), and one section was mounted on a glass slide, and the second on film (P.A.L.M.), which had been fixed onto a glass slide (Micro Slides Colorfrost, VWR Scientific, Media, PA). The section mounted on a glass slide was after fixed in cold acetone, and stained with Mayer's Haematoxylin (Sigma, St. Louis, MO). A pathologist analyzed the samples for diagnosis and grade. The clinical stage was estimated from the accompanying surgical pathology and clinical reports to verify the Dukes classification. The section mounted on film was after fixed for five minutes in 100% ethanol, counter stained for 1 minute in eosin/100% ethanol (100μg of Eosin in 100ml of dehydrated ethanol), quickly soaked once in 100% ethanol to remove the free stain, and air dried for 10 minutes.

Before use in LCM, the membrane (LPC-MEMBRANE PEN FOIL 1.35 µm No 8100, P.A.L.M. GmbH Mikrolaser Technologie, Bernried, Germany) and slides were pretreated to abolish RNases, and to enhance the attachment of the tissue sample onto the film. Briefly, the slides were washed in DEP H_2O, and the film was washed in RNase AWAY (Molecular Bioproducts, Inc., San Diego, CA) and rinsed in DEP H_2O. After attaching the film onto the glass slides, the slides were baked at +120°C for 8 hours, treated with TI-SAD (Diagnostic Products Corporation, Los Angeles, CA, 1:50 in DEP H_2O, filtered through cotton wool), and incubated at +37°C for 30 minutes. Immediately before use, a 10µl aliquot of RNase inhibitor solution (Rnasin Inhibitor 2500U=33U/µl N211A, Promega GmbH, Mannheim, Germany, 0.5µl in 400µl of freezing solution, containing 0.15 mol NaCl, 10 mmol Tris pH 8.0, 0.25 mmol dithiothreitol) was spread onto the film, where the tissue sample was to be mounted.

The tissue sections mounted on film were used for LCM. Approximately 2000 epithelial cells/sample were captured using the PALM Robot-Microbeam technology (P.A.L.M. Mikrolaser Technologie, Carl Zeiss, Inc., Thornwood, NY), coupled into Zeiss Axiovert 135 microscope (Carl Zeiss Jena GmbH, Jena, Germany). The surrounding stroma in the normal mucosa, and the occasional intervening stromal components in cancer samples, were included. The captured cells were put in tubes in 100% ethanol and preserved at -80°C.

Example 2- RNA Extraction and Amplification.

Zymo-Spin Column (Zymo Research, Orange, CA 92867) was used to extract total RNA from the LCM captured samples. About 2 ng of total RNA was resuspended in 10 ul of water and 2 rounds of the T7 RNA polymerase based amplification were performed to yield about 50 ug of amplified RNA.

Example 3- DNA Microarray Hybridization and Quantitation.

A set of DNA microarrays consisting of approximately 23,000 human DNA clones was used to test the samples by use of the humanU133a chip obtained and commercially available from Affymetrix, Inc. Total RNA obtained and prepared as outlined above and applied to the chips and analyzed by Agilent BioAnalyzer according to the manufacturer's protocol. All 63 samples passed the quality control standards and the data were used for marker selection.

Chip intensity data was analyzed using MAS Version 5.0 software commercially available from Affymetrix, Inc. ("MAS 5.0"). An unsupervised analysis was used to identify two genes that distinguish patients that would relapse from those who would not as follows.

The chip intensity data obtained as described was the input for the unsupervised clustering software commercially available as PARTEK version 5.1 software. This unsupervised clustering algorithm identified a group of 20 patients with a high frequency of relapse (13 relapsers and 7 survivors). From the original 23,000 genes, t-testing analysis selected 276 genes that significantly differentially expressed in these patients. From this group, two genes were selected that best distinguish relapsing patients from those that do not relapse: Human intestinal peptide-associated transporter (Seq. ID. No. 3) and Homo sapiens fatty acid binding protein 1 (Seq. ID No. 1). These two genes are down-regulated (in fact, they are turned off or not expressed) in the relapsing patients from this patient group.

Supervised analysis was then conducted to further discriminate relapsing patients from those who did not relapse in the remaining 43 patients. This group of patient data was then divided into the following groups: 27 patients were assigned as the training set and 16 patients were assigned as the testing set. This ensured that the same data was not used to both identify markers and then validate their utility.

An unequal variance t-test was performed on the training set. From a list of 28 genes that have significant corrected p values, MHC II-DR-B was chosen. These genes are down-regulated in relapsers. MHC II-DR-B (Seq. ID No. 2) also had the smallest p-value.

In an additional round of supervised analysis, a variable selection procedure for linear discriminant analysis was implemented using the Partek Version 5.0 software described above to separate relapsers from survivors in the training set. The search method was forward selection. The variable selected with the lowest posterior error was immunoglobulin-like transcript 5 protein (Seq. ID No. 4). A Cox proportional hazard model (using "S Plus" software from Insightful, Inc.) was then used for gene selection to confirm gene selection identified above for survival time. In each cycle of total 27 cycles, each of the 27 patients in the training set was held out, the remaining 26 patients were used in the univariate Cox model regression to assess the strength of association of gene expression with the patient survival time. The strength of such association was evaluated by the corresponding estimated standardized parameter estimate and P value returned from the Cox model regression. P value of 0.01 was used as the threshold to select top genes from each cycle of the leave-one-out gene selection. The top genes selected from each cycle were then compared in order to select those genes that showed up in at least 26 times in the total of 27 leave-one-out gene selection cycles. A total of 70 genes were selected and both MHC II-DR-B and immunoglobulin-like transcript 5 protein were among them (Again, showing down regulation).

Construction of a multiple-gene predictor: Two genes, MHC II-DR-B and immunoglobulin-like transcript 5 protein were used to produce a predictor using linear discriminant analysis. The voting score was defined as the posterior probability of relapse. If the patient score was greater than 0.5, the patient was classified as a relapser. If the patient score was less than 0.5, the patient was classified as a survivor. The predictor was tested on the training set.

Cross-validation and evaluation of predictor: Performance of the predictor should be determined on an independent data set because most classification methods work well on the examples that were used in their establishment. The 16 patients test set was used to assess prediction accuracy. The cutoff for the classification was determined by using a ROC curve. With the selected cutoff, the numbers of correct prediction for relapse and survival patients in the test set were determined.

Overall prediction: Gene expression profiling of 63 Duke's B colon cancer patients led to identification of 4 genes that have differential expression (down regulation or turned off) in these patients. These genes are Seq. ID No. 1, Seq. ID No. 2, Seq. ID No. 3, and Seq. ID No. 4. Thirty-six of the patients have remained disease-free for more than 3 years while 27 patients had tumor relapse within 3 years. Using the 3 gene markers portfolio of Seq. ID No. 2 , Seq. ID No. 3, and Seq. ID No. 4, 22 of the 27 relapse patients and 27 of 36 disease-free patients are identified correctly. This result represents a sensitivity of 82% and a specificity of 75%. The positive predictive value is 71% and the negative predictive value is 84%.

Example 4: Further Sampling

Frozen tumor specimens from 74 coded Dukes' B colon cancer patients were then studied. Primary tumor and adjacent non-neoplastic colon tissue were collected at the time of surgery. The histopathology of each specimen was reviewed to confirm diagnosis and uniform involvement with tumor. Regions chosen for analysis contained a tumor cellularity greater than 50% with no mixed histology. Uniform follow-up information was also available.

Example 5: Gene Expression Analysis

Total RNA was extracted from the samples of Example 4 according to the method described in Examples 1-3. Arrays were scanned using standard Affymetrix protocols and scanners. For subsequent analysis, each probe set was considered as a separate gene. Expression values for each gene were calculated by using Affymetrix GeneChip analysis software MAS 5.0. All data used for subsequent analysis passed quality control criteria.

Statistical Methods

Gene expression data were first subjected to a variation filter that excluded genes called "absent" in all the samples. Of the 22,000 genes considered, 17,616 passed this filter and were used for clustering. Prior to the hierarchical clustering, each gene was divided by its median expression level in the patients. Genes that showed greater than 4-fold changes over the mean expression level in at least 10% of the patients were included in the clustering. To identify patient subgroups with distinct genetic profiles, average linkage hierarchical clustering and k-mean clustering was performed by using GeneSpring 5.0 (San Jose, CA) and Partek 5.1 software (St. Louis, MO), respectively. T-tests with Bonferroni corrections were used to identify genes that have different expression levels between 2 patient subgroups implicated by the clustering result. A Bonferroni corrected P value of 0.01 was chosen as the threshold for gene selection. Patients in each cluster that had a distinct expression profile were further examined with the outcome information.

In order to identify gene markers that can discriminate the relapse and the disease-free patients, each subgroup of the patients was analyzed separately as described further below. All the statistical analyses were performed using S-Plus software (Insightful, VA).

Patient and Tumor Characteristics

Clinical and pathological features of the patients and their tumors are summarized in Table 1. The patients had information on age, gender, TNM stage, grade, tumor size and tumor location. Seventy-three of the 74 patients had data on the number of lymph nodes that were examined, and 72 of the 74 patients had estimated tumor size information. The patient and tumor characteristics did not differ significantly between the relapse and non-relapse patients. None of the patients received pre-operative treatment. A minimum of 3 years of follow-up data was available for all the patients in the study.

Patient Subgroups Identified by Genetic Profiles

Unsupervised hierarchical clustering analysis resulted in a cluster of the 74 patients on the basis of the similarities of their expression profiles measured over 17,000 significant genes. Two subgroups of patients were identified that have over 600 differentially expressed genes between them ($p < 0.00001$). The larger subgroup and the smaller subgroup contained 54 and 20 patients, respectively. In the larger subgroup of the 54 patients only 18 patients (33%) developed tumor relapse within 3 years whereas in the smaller subgroup of the 20 patients 13 patients (65%) had progressive diseases. Chi square analysis gave a p value of 0.028.

Two dominant gene clusters that had drastic differential expression between the two types of tumors were selected and examined. The first gene cluster had a group of down-regulated genes in the smaller subgroup of the 20 patients, represented by liver-intestine specific cadherin 17, fatty acid binding protein 1, caudal type homeo box transcription factors CDX1 and CDX2, mucin and cadherin-like protein MUCDHL. The second gene cluster is represented by a group of up-regulated genes in the smaller subgroup including serum-inducible kinase SNK, annexin A1, B cell RAG associated protein, calbindin 2, and tumor antigen L6. The smaller subgroup of the 20 patients thus represent less differentiated tumors on the basis of their genetic profiles.

Gene Signature and its Prognostic Value

In order to identify gene markers that can discriminate the relapse and the disease-free patients, each subgroup of the patients were analyzed separately. The patients in each subgroup were first divided into a training set and a testing set with approximately equal number of patients. The training set was used to select the gene markers and to build a prognostic signature. The testing set was used for independent validation. In the larger subgroup of the 54 tumors, 36 patients had remained disease-free for at least 3 years after their initial diagnosis and 18 patients had developed tumor relapse with 3 years. The 54 patients were divided into two groups. The training set contained 21 disease-free patients and 6 relapse patients. In the smaller subgroup of the 20 tumors, 7 patients had remained disease-free for at least 3 years and 13 patients had developed tumor relapse with 3 years. The 20 patients were divided into two groups. The training set contained 4 disease-free patients and 7 relapse patients. To identify a gene signature that discriminates the good prognosis group from the poor prognosis group, a supervised classification method was used on each of the training sets. Univariate Cox proportional hazards regression was used to identify genes whose expression levels are correlated to patient survival time. Genes were selected using p-values less than 0.02 as the selection criteria. Next, t-tests were performed on the selected genes to determine the significance of the differential expression between relapse and disease-free patients ($P < 0.01$). To avoid selection of genes that over-fit the training set, re-sampling of 100 times was performed with the t-test in order to search for genes that have significant p values in more than 80% of the re-sampling tests. Seven genes (Table 2) were selected from the 27 patient training set and 15 genes (Table 3) were selected from the 11 patient training set. Taking the 22 genes and cadherin 17 together, a Cox model to predict patient recurrence was built using the S-Plus software. The Kaplan-Meier survival analysis showed a clear difference in the probability that patients would remain disease free between the group predicted with good prognosis and the group predicted with poor prognosis (Fig. 3).

Several genes are related to cell proliferation or tumor progression. For example, tyrosine 3 monooxygenase tryptophan 5-monooxygenase activation protein (YWHAH) belongs to 14-3-3 family of proteins that is responsible for G2 cell cycle control in response to DNA damage in human cells. RCC1 is another cell cycle gene involved in the regulation of onset of chromosome condensation. BTEB2 is a zinc finger transcription factor that has been implicated as a beta-catenin independent Wnt-1 responsive genes. A few genes are likely involved in local immune responses. Immunoglobulin-like transcript 5 protein is a common inhibitory receptor for MHC I molecules. A unique member of the gelsolin/villin family capping protein, CAPG is primarily expressed in macrophages. LAT is a highly tyrosine phosphorylated protein that links T cell receptor to cellular activation. Thus both tumor cell- and immune cell-expressed genes can be used as prognostic factors for patient recurrence.

In order to validate the 23-gene prognostic signature, the patients in the two testing sets that included 27 patients from the larger subgroup and 9 patients from the smaller subgroup were combined and outcome was predicted for the 36 independent patients in the testing sets. This testing set consisted of 18 patients who developed tumor relapses within 3 years and 18 patients who had remained disease free for more than 3 years. The prediction resulted in 13 correct relapse classification and 15 correct disease-free classifications. The overall performance accuracy was 78% (28 of 36) with a sensitivity of 72% (13 of 18) and a specificity of 83% (15 of 18). This performance indicates that the Dukes' B patients that have a value below the threshold of the prognostic signature have a 13-fold odds ratio of (95% CI: 2.6, 65; p=0.003) developing a tumor relapse within 3 years compared with those that have a value above the threshold of the prognostic signature. Furthermore, the Kaplan-Meier survival analysis showed a significant difference in the probability that patients would remain disease free between the group predicted with good prognosis and the group predicted with poor prognosis (P < 0.0001). In a multivariate Cox proportional hazards regression, the estimated hazards ratio for tumor recurrence was 0.41 (95% confidence interval, 0.24 to 0.71; P = 0.001), indicating that the 23-gene set represents a prognosis signature and it is inversely associated with a higher risk of tumor recurrence. Using the seven gene portfolio (Table 2), an 83% sensitivity and 80% specificity were obtained (based on a

12 relapse and 15 survivor sample set). Using the 15 gene portfolio (Table 3), a 50% sensitivity and 100% specificity were obtained (based on 6 relapse and three survivor sample sets). Figures 1 and 2 are graphical portrayals of the Kaplan-Meier analyses for the seven and fifteen gene portfolios respectively.

Furthermore, as these results demonstrate, prognosis can be derived from gene expression profiles of the primary tumor.

Table 1. Clinical and Pathological Characteristics of Patients and Their Tumors

Characteristics	Disease-free		Recurrence		P Value*
	no. of patients (%)				
Age	43		31		0.7649
Mean	58.93		58.06		
Sex	43		31		0.8778
Female	23	(53)	18	(58)	
Male	20	(47)	13	(42)	
T Stage	43		31		0.2035
2	12	(28)	5	(16)	
3	29	(67)	26	(84)	
4	2	(5)	0	(0)	
Differentiation	43		31		0.4082
Poor	5	(12)	6	(19)	
Moderate	37	(86)	23	(74)	
Well	1	(2)	2	(6)	
Tumor size	41		31		0.1575
<5	29	(71)	16	(52)	
>=5	12	(29)	15	(48)	
Location	43		31		0.7997
LC	1	(2)	1	(3)	
RC	17	(40)	10	(32)	
TC	6	(14)	3	(10)	
SC	19	(44)	17	(55)	
Number of LN examined	43		30		0.0456
Mean	12.81		8.63		

* P values for Age, Lymph node number and Tumor content are obtained by t tests; P values for others are obtained by χ^2 tests.

Table 2: 7 Gene List

Accession	Seq. I.D No.
AF009643.1	7
NM_003405.1	8
X06130.1	9
AB030824.1	10
NM_001747.1	11
AF036906.1	12
BC005286.1	13

Table 3: 15 Gene List

Accession	Seq. I.D. No.
NM_012345.1	14
NM_030955.1	15
NM_001474.1	16
AF239764.1	17
D13368.1	18
NM_012387.1	19
NM_016611.1	20
NM_014792.1	21
NM_017937.1	22
NM_001645.2	23
AL545035	24
NM_022078.1	25
AL133089.1	26
NM_001271.1	27
AL137428.1	28

Table 4. Twenty-three genes form the prognostic signature.

Direction of change	Seq. ID No.	P value (Cox)	Gene Description
-	7	0.0011	immunoglobulin-like transcript 5 protein
-	8	0.0016	tyrosine 3-monooxygenasetryptophan 5-monooxygenase activation protein
-	9	0.0024	cell cycle gene RCC1
+	10	0.0027	transcription factor BTEB2
-	11	0.0045	capping protein (actin filament), gelsolin-like (CAPG)
-	12	0.0012	linker for activation of T cells (LAT)
-	13	0.0046	Lafora disease (laforin)
-	14	0.0110	nuclear fragile X mental retardation protein interacting protein 1 (NUFIP1)
+	15	0.0126	disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 12 (ADAMTS12)
+	16	0.0126	G antigen 4 (GAGE4)
+	17	0.0130	EGF-like module-containing mucin-like receptor EMR3
+	18	0.0131	alanine:glyoxylate aminotransferase
+	19	0.0131	peptidyl arginine deiminase, type V (PAD)
+	20	0.0136	potassium inwardly-rectifying channel, subfamily K, member 4 (KCNK4)
+	21	0.0139	KIAA0125 gene product (KIAA0125)
+	22	0.0142	hypothetical protein FLJ20712 (FLJ20712)
+	23	0.0145	apolipoprotein C-I (APOC1)
+	24	0.0146	Consensus includes gb:AL545035
+	25	0.0149	hypothetical protein FLJ12455 (FLJ12455)
+	26	0.0150	Consensus includes gb:AL133089.1
+	27	0.0151	chromodomain helicase DNA binding protein 2 (CHD2)
+	28	0.0152	Consensus includes gb:AL137428.1
N/A	6	Not tested	Cadherin 17